Avoiding the drunkard's search

Avoiding the Drunkard's Search: Investigating Collection Strategies for Building a Twitter Dataset

Clare Llewellyn
University of Edinburgh
2F2 Buccleuch Place
Edinburgh, UK
s1053147@sms.ed.ac.uk

Laura Cram
University of Edinburgh
2F2 Buccleuch Place
Edinburgh, UK
Laura.Cram@ed.ac.uk

Adrian Favero
University of Edinburgh
2F2 Buccleuch Place
Edinburgh, UK
A.Favero@ed.ac.uk

ABSTRACT

We investigate methods for collecting data to form an archive on the debate within Twitter surrounding the UK's inclusion in the EU. We use three strategies, gathering data using hashtags, extracting data from the random stream and collecting from users known to be discussing the debate. We explore the various bias in the resulting datasets.

Keywords

Data Analytics; Social Media Analysis; Data Selection

1. INTRODUCTION

We are gathering data from Twitter to create an archive on the debate surrounding the UK's inclusion in the European Union (EU). We are tracking opinion leading up to a referendum on the UK's membership. Twitter is being used to find out what people are saying and to investigate how this changes over time. Twitter can be used to track trends in response to emerging events and this analysis allows us to gain a more nuanced understanding of those who are motivated to comment on UK-EU-related topics.

Twitter studies are often criticised because they employ a 'drunkard's search' method, where researchers only look at what is easy to find, like a drunk person looking for keys under a street light because that is the only place where they can see.

An easy way to generate a topic specific Twitter dataset is by querying the Twitter API using hashtags. This method provides data that has been annotated by authors using a keyword or phrase that generally suggests a topic label or a context. The generation of a dataset using this method, however, biases the content in favour of the hashtags chosen. Badly chosen hashtags will mean not all data is covered, hashtags may change over time as debate evolves [4] and data maybe missed if it is not marked with a hashtag.

To address this problem this paper contrasts three methods for collecting data from Twitter: 1) using hashtags chosen by an expert panel as search queries; 2) collecting the random sample without specified search terms and extracting appropriate data [2]; 3) collecting from specific users that are known to be contributing to the debate [3].

JCDL '16 June 19-23, 2016, Newark, NJ, USA

ACM ISBN 978-1-4503-4229-2/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2910896.2925433

2. BACKGROUND

Twitter provides access to a small sample of data through two API methods, a streaming method and a search method. Both give access to 'a small sampling' [1] of the data as it is produced (streaming) or that previously published (search), as results from a query or a random sample. It is possible to share datasets by providing the user id, tweet id and software for gathering data directly from Twitter.

UK citizens will vote on whether to remain within the EU in a referendum that is to take place on the June 23, 2016. The debate over whether the UK should remain as part of the EU is between those who favour remaining as a member (pro-remain) and those who wish the UK to leave the EU (pro-leave). We are monitoring how shifts in opinion relate to the wider public debate and the extent to which Twitter can be used to measure public opinion in relation to the EU.

Data has been gathered as part of an on-going process since Aug 6, 2015 using three strategies:

The Hashtags Set data is collected from the streaming API using UK-EU specific hashtags, chosen by a panel of experts, as query terms. This includes referendum specific terms such as #brexit, #euref, those reflecting topics which will likely be debated such as #migrants and #refugees and more general relevant terms such as #EU and #Europe.

The Stream Set data is extracted from the streaming API using a method based on [2]. This involves collecting the data from the streaming API without any search terms, thereby receiving a random selection. Data is then extracted from this selection using a set of commonly used relevant terms. This gives a topic specific set from the random set. This topic specific set is analysed and the top 100 unigram, bigram and trigram terms are identified. Two annotators then assign each of these terms as relevant or not to UK-EU discussion and the relevant terms are used to search the wider random set to expand the topic specific set. This approach aims to reduce the bias introduced through human defined search terms.

The Official Set data is collected from a group of users that are known to be discussing the referendum, the official campaign groups, StrongerIn, LeaveEUOfficial, Grassroots_Out and Vote_Leave. The data is collected daily via the search API.

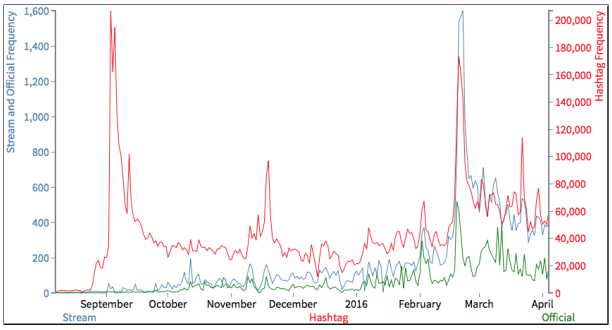

Figure 1: The Frequency of tweets over time.

3. ANALYSIS

We present results based on a comparison of the three data collection methods. We analyse the overall frequency of tweets, the frequency over time in response to specific events, an evaluation of relevance to topic, and an evaluation of how hashtags are used in each set.

During an eight-month collection period the hashtag set contained over ten million tweets, the stream set over forty two thousand and the official set over sixteen thousand. Showing that the hashtag approach collects the largest set by a considerable amount. The frequency of tweets over time graph (Fig.1) shows various similar peaks in the data indicating that all collection strategies are picking up an increase on specific dates. There is a peak in data collected on October 12 2015 when the StrongerIn campaign was launched, and on October 9 2015 after a speech on the EU by David Cameron. There is a large amount of data collected by hashtags on September 3 2015 that is not in the other sets, this was found to be related to refugees and migrants.

Relevance is evaluated independently by three annotators in two tasks. 100 tweets were randomly selected from each dataset. Firstly the annotators were asked to determine if each tweet was directly relevant to the debate on the UK-EU referendum. In the second task the annotators were asked to determine if each tweet was referendum relevant or about a topic that would likely influence voter opinion.

We can see that data in the official and the stream set is more relevant to both the referendum and topics relating to the referendum than the hashtag set (Tab. 1). The hashtag has a low relevance score for 'directly relevant to the referendum debate' but this rises significantly when the topics that will influence the debate are considered. The results indicate that although the hashtag set contains non-relevant information it also covers the topics likely to influence voters not identified in the other sets.

We investigated if hashtags are used differently in the collections through the use of three specific hashtags, one pro-remain (#strongerin) one pro-leave (#leaveeu) and one neutral (#brexit). For each of the sets we gathered 50 random tweets that contained each of the hashtags. Two annotators were asked to mark if each tweet was pro-remain, pro-leave or neutral. We can see (Tab. 3) that all three of the hashtags, thought to represent pro-leave, pro-remain and neutral points of view, are used in tweets that have a pro-leave sentiment. Although #strongerin is used to present a pro-remain opinion in the official set by the pro-remain campaign group.

Table 1: Relevance of data to the EU

	Task 1				Task 2			
	A1	A2	A3	Average	A1	A2	A3	Average
Hashtag	18	8	24	16.67	49	38	68	51.67
Official	91	72	85	82.67	94	80	95	89.67
Stream	95	58	83	78.67	96	79	92	89

Table 2: Opinion (Leave/Remain/Neutral)

		leaveeu %			strongerin %			brexit %		
		L	R	N	L	R	N	L	R	N
Hashtag	A1	94	0	6	42	46	12	76	10	16
	A2	84	0	16	52	42	6	58	6	34
Official	A1	94	0	6	0	100	0	96	0	4
	A2	80	0	20	2	96	2	88	0	12
Stream	A1	100	0	0	60	32	8	78	14	8
	A2	88	0	12	62	34	6	48	4	46

4. CONCLUSIONS

Both the stream and hashtags sets are heavily influenced by the terms used for data collection. The terms differ when automatically extracted (the stream set) or chosen by experts (the hashtag set). The automatic method is most similar to the official set in terms of relevance and frequency of tweets over time. These sets are both very specific to the topic and small in comparison to the hashtag set. The expert method includes a variety of terms that the experts expect will become discussion topics over the longer-term referendum debate. This approach therefore, has a low direct relevance but it gathers information on wider associated topics likely to influence voter choices that may be missed by the other two methods. We cannot extrapolate from this set that these topics will influence the debate, only that they are being discussed. The top hashtag lists for each set and the use of #brexit, #strongerin and #leaveeu suggest that either all of the data selection strategies to collect the data are biased towards the pro-leave opinion or that the data from Twitter contains a strong pro-leave opinion. It is also likely that the term brexit is not as neutral as we thought. Future work includes, gathering relevant data using a supervised machine learning approach, using frequent hashtags in the official/stream sets to update the query terms in the hashtag set and comparing the content of tweets that contain hashtags and those that do not.

5. REFERENCES

[1] Twitter developer pages. https://support.twitter.com/articles/160385. Accessed: 2016-01-22.

[2] C. Llewellyn, C. Grover, B. Alex, J. Oberlander, and R. Tobin. Extracting a topic specific dataset from a twitter archive. In *Research and Advanced Technology for Digital Libraries*, pages 364–367. Springer, 2015.

[3] D. O'Callaghan, N. Prucha, D. Greene, M. Conway, J. Carthy, and P. Cunningham. Online social media in the syria conflict: Encompassing the extremes and the in-betweens. In *ASONAM, 2014 IEEE/ACM*, pages 409–416. IEEE, 2014.

[4] Z. Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400*, 2014.